User-centric evaluation of recommender systems in social learning platforms: Accuracy is just the tip of the iceberg

Accuracy is just the tip of the iceberg: A Data-centric vs. User-centric Evaluation

Soude Fazeli, Hendrik Drachsler, Marlies Bitter-Rijpkema, Francis Brouns, Wim van der Vegt, and Peter B. Sloep

Abstract—Recommender systems provide users with content they might be interested in. Conventionally, recommender systems are evaluated mostly by using prediction accuracy metrics only. But the ultimate goal of a recommender system is to increase user satisfaction. Therefore, evaluations that measure user satisfaction should be also performed before deploying a recommender system to a real target environment. Such evaluations are laborious and complicated compared to the traditional, data-centric evaluations, though. In this study, we investigate the added value of user-centric evaluations and how user satisfaction of a recommender system is related to its performance in terms of accuracy metrics. We conduct both a data-centric evaluation and a user-centric evaluation on the same data collected from an authentic social learning platform. Our findings suggest that user-centric evaluation results are not necessarily in line with data-centric evaluation results. We conclude that the traditional evaluation of recommender systems in terms of prediction accuracy does not suffice to judge performance of recommender systems on the user side. Moreover, the user-centric evaluation provides valuable insights on how candidate algorithms perform on each of the five quality metrics: usefulness, accuracy, novelty, diversity, and serendipity of the recommendations.

Index Terms—recommender systems, evaluation, social, learning, accuracy, performance

1 Introduction

Recommender systems provide a user with the content she or he might be interested in. They have become increasingly popular because of their successful applications in the e-commerce field, such as with Amazon and eBay. Recommender systems have been introduced in the educational domain as a practical solution to help users find suitable content that can support their learning process [1], [2]. Traditionally, recommender systems have been evaluated according to accuracy metrics in the Information Retrieval area. However, such evaluations do not answer the question whether the users are really satisfied with the recommendations as indicated by the accuracy metrics. Recently, researchers have realized that the goal of a recommender system goes beyond the accuracy metrics [3], [4]. This has prompted two major changes in the field of recommender systems. The first change, indicated by McNee et al. [4], is that "being accurate is not enough". These authors also emphasized that researchers should "study recommenders from a user-centric perspective to make them not only accurate and helpful, but also a pleasure to use" [4]. The second change has been introduced as "a broadening of the scope of research regarding the system aspects to investigate beyond just the algorithm of the recommender" [3], [5]. Following this, McNee et al. suggest researchers to also study the aspects of "Human-Recommender Interaction" [6]. Martin [7] claimed in his keynote to the ACM RecSys 2009 conference that around 50% of a recommender's commercial success goes to the aspects of "Human-Recommender Interaction" while the algorithm matters for 5% only (Martin2009).

The importance of the user perspective has been realized even more in the educational domain [1], [8], [9]. Indeed, the main goal of the educational recommender systems extends well beyond accurate predictions and should also take into account quality metrics such as usefulness, novelty, or diversity of the recommendations.

Although the importance of user-centric evaluations has become quite clear and vital, the majority of recommender system studies still solely report the traditional, data-centric evaluation results. Many of them are based on some implicit feedback like Click Through Rate (CTR) [10], [11], which hardly reflect users' satisfaction and their perceived usefulness on the recommendations made for them. However, traditional offline user-centric evaluations, such as those based on CTR, are more straightforward to conduct compared to user-centric evaluations based on explicit questionnaires. There are several reasons that make user-centric evaluations complicated to carry out. First, they can easily fail due to the lack of a sufficient numbers of participants. Second, it is also quite tricky to design an experimental protocol such that it attracts users instead of detracting them. The users' task should be defined clearly and simply, helping users to spend a fair amount of time on the task and also making sure not to be misunderstood. Third, setting up a test bed as an experimental environment is a time-consuming and delicate job. Fourth, user-centric evaluations can take up to several months and they are quite vulnerable to the availability and loading speed of the experimental environment (a social platform in this

- *S. Fazeli is a postdoctoral researcher at Delft University of Technology, Delft, the Netherlands. E-mail: s.fazeli@tudelft.nl*
- *Hendrik Drachsler, Marlies Bitter-Rijpkema, Francis Brouns, Wim van der Vegt, and Peter B. Sloep are with the Open University of the Netherlands. E-mail: {hendrik.drachsler,marlies.bitter,francis.brouns,wim.vandervegt, peter.sloep}@ou.nl.*

Manuscript received April 19, 2005; revised August 26, 2015.

study); continuous availability of the participants is also a concern. Moreover, many user-centric evaluations are conducted using crowdsourcing. Although that is a valid approach, it has its limitations [9], [12]. In crowdsourcing, tasks, reliability and accuracy of the collected feedback data is sometimes questionable since there are of course differences between "cheap labor" workers and expensive experts [13].

In this study, we want to investigate what the added value of user-centric evaluations is precisely because of the complexity of carrying them out. There is no point in conducting them if they turn out to be less useful than anticipated. So our main research question is:

RQ: In social learning platforms, how is user satisfaction with recommender systems related to the performance of such systems measured in terms of their accuracy?

We conduct both a traditional, data-centric evaluation and a user-centric evaluation. Such an evaluation aims to answer our research question by using both a proposed graph-based approach and two state-of-the-art recommender algorithms within an authentic social learning platform developed by the eContentPlus Open Discovery Space (ODS) project (http://opendiscoveryspace.eu). By the term social learning platform, we refer to those platforms that combine traditional learning management systems (LMS) with commercial social networks like Facebook to provide easy content creation, access, sharing, bookmarking, etc. Beside forums and chat communities often provided in standard-LMSs, they let users establish more connections and improve their networks of peers.

The rest of the paper is structured as follows: In Section 2, we describe the experimental method used including the algorithms, the data, and the evaluation settings. Section 3 presents the experimental results including results of both traditional evaluation and user-centric evaluation. Section 4 discusses the extent to which the results answer the research question defined in this study, and finally, draws conclusions.

2 Experimental Method

To address the main research question, we run two sets of evaluations: 1. A conventional evaluation study for comparing performance of recommender systems based on traditional accuracy metrics, and 2. A user-centric evaluation as an online study to ask the actual users for their feedback on the recommendations made for them.

In this section, we first provide a description of the data used. Second, we give an overview of the recommender algorithms chosen for this study. Finally, we explain the settings of both evaluation methods (data-centric and user-centric).

2.1 Data

The data used in this study comes from the Open Discovery Space (ODS) platform. According to the official website, "[The] Open Discovery Space [project] addresses the challenge of modernizing school education by engaging teachers, students, parents and policymakers in a first of its kind effort to create a pan-European eLearning environment to promote more flexible and creative ways of learning by improving the way educational content is produced, accessed and used" (http://opendiscoveryspace.eu). The platform is the online area where all the ODS stakeholders meet.

The ODS data, collected through the platform, contains social data of users such as ratings, tags, reviews, etc. on learning resources, communities, groups, etc. The ODS data complies with the CAM (Context Automated Metadata) format [14], which provides a standard metadata specification for collecting and storing social data. A CAM schema aims to store whatever has attracted users' attention while the users are working with the platform. It also stores users' interaction with the platform such as rating, tagging, etc. A CAM schema records an event and its details when a user performs an action within a platform. The metadata stored in the CAM format describe all types of users' feedback and, therefore, can be further converted to the input data required for making recommendations for the users.

The ODS dataset contains interaction data (9117 events) of 2,567 users with 3,392 objects. It should be noted that the data is too sparse in terms of user transactions (degree of sparsity=99.86%) to make recommendations with classical recommender systems. Our experiment happens in the learning domain in which datasets are generally smaller than the ones in e-commerce [3]. Although the dataset is rather small, it realistically represents the current ODS platform. Sparsity often occurs in educational settings and requires specific adjustments to the recommendation approach as shown by [15]. We therefore took sparsity into account as one of the properties of the data when selecting the most appropriate algorithms for ODS. The data span the time period from May 2013 until October 2015.

2.2 Algorithms

The first step in developing a recommender system is to find out with what kind of input data to fuel the recommender engine. The items in the ODS platform are learning resources, communities, groups, and discussion posts. The user activities in the ODS platform is mainly implicit user feedback coming from tracking data, such as viewing, bookmarking, downloading a resource or joining a community. Therefore, Collaborative Filtering recommenders can be applied. Collaborative Filtering (CF) methods make recommendations for a target user based on other users' opinions and interests [16], [17]. Content-based methods should be used when there is no user rating information (5-star, binary, unary) available. However, as is also indicated in recommender systems studies [18], "even if very few ratings are available, simple rating-based predictors outperform purely metadata-based ones". This is likely to be due to the large difference between the item descriptions and the items themselves. And users rate items, not their descriptions. In general, the CF algorithms are categorized according to their *type* and *technique*. *Type* refers to model-based and memory-based algorithms and *technique* refers to user-based and item-based algorithms. In the rest of this section, we are going to describe different *types* and *techniques* from the CF family. In this study, we try to make use of the algorithms from all categories as well as a graph-based method we proposed in our previous work [15].

2.2.1 Memory-based recommender systems

Most of the CF algorithms are based on k-nearest-neighbour (kNN) methods (k being the size of the neighbourhood). They have proven to be quite successful [19]. kNN tries to find like-minded users and introduces them as the nearest neighbours of a target user for whom recommendations are generated. The kNN algorithms create a graph of users in which nodes are users and the edges are similarity relations between them. Depending on whether the data includes explicit user feedback (e.g. 5-star ratings) or implicit user feedback (e.g. views, downloads, clicks, etc.), different similarity measures are appropriate. The data used in this study provides implicit user feedback in the form of (userID,itemID) tuples, with, "item" referring to learning objects, communities, groups, etc. in the ODS platform. This kind of data is also known as "positive feedback only" since they present only users' interests in items where there is no negative feedback expressed by users on items [20]. Therefore, some of the similarity measures such as Pearson correlation are not suitable because they require explicit user feedback i.e. in forms of 5-star ratings explicitly expressed by users. As one of the popular similarity measures, we used the Jaccard coefficient since the data includes implicit user feedback in binary format [21]. In this study, we use both user-based and item-based CF algorithms since we make use of users' interactions and activities. User-based algorithms try to find patterns of similarity between users in order to make recommendations; item-based algorithms follow the same process but are based on similarity between items.

2.2.2 A graph-based recommender system

Although the kNN methods are quite popular in the recommender systems area, they have two shortcomings. First, they usually do not work well when the user feedback data is sparse, which is often the case in the educational domain [1]. Second, they are only limited to k neighbours for each user. Thus two users who have not shown an interest in a common set of items cannot be connected, even though they might be a good source of information for each other. Therefore, the implicit user networks inferred by these methods are always affected by this constraint, which in turn may affect the process of knowledge sharing and peer collaborations in online learning platforms. Note that platforms such as ODS have been set up exactly to foster peer collaboration, learning from each other, and other activities that promote the shared construction of knowledge. To address the sparsity issue and the restriction to k neighbours only, we employed a graph-based approach [15], [22]. Such an approach extends and improves the kNN's process of finding neighbours, by invoking graph search algorithms. The graph-based approach first forms a graph in which nodes are users and edges are similarity relations between users. Then, it collects recommendations for a target user by "walking" through the target user's neighbours. The graph-based approach is memory-based and user-based. Approaches to improve performance of recommenders by using graph-walking algorithms do exist already and report positive effects in different domains [22]–[25]. However, almost all use data regarding either social relations between users or inter-user trust relations; these are not available for the datasets used in this study. Indeed, we use the graph-based approach with the aim of supporting the target users of social learning platforms to identify their potentially interesting and novel neighbours.

2.2.3 Model-based recommender systems

Model-based methods create models of users' preferences using probabilistic approaches such as neural networks, Bayesian networks, and algebraic approaches such as those using eigenvectors. They are known for their fast performance as they create users' preferences models offline but they need a full set of users' preferences to develop a user model. Moreover, model-based methods often prove to be costly in terms of required resources and maintenance efforts. In this study, we need model-based CFs that can deal with implicit data feedback. Rendle et al. [26] applied their Bayesian Personalized Ranking (BPR) to the state-of-the-art matrix factorization models to improve the learning process in the Bayesian model used (BPRMF). We choose the BPRMF for our experimental study since it can work well with the data we used.

2.3 Data-centric Evaluation

We ran a data-centric evaluation to assess performance of the candidate recommender algorithms in terms of accuracy metrics in the Information Retrieval area. Within this conventional type of studies, there is no direct interaction with the actual users. The algorithms are measured according to, precision and recall to measure the accuracy of the recommendations generated [27]. Precision is defined as the percentage of recommended items that are relevant to the user (i.e., ratio of the number of items recommended that were relevant to the total number of recommended items). Recall shows the probability that a relevant item is recommended (i.e., the number of items recommended that were relevant divided by the total number of relevant items in the entire test set). Both precision and recall range from 0 to 1. In this study, 80% of the data was randomly selected and assigned to the training set and the rest was considered the test set. These metrics and settings are commonly used for empirical studies on recommender systems [27].

2.4 User-centric Evaluation

We conducted a user-centric evaluation to measure the perceived quality of the recommendations made for ODS users. The user-centric evaluation consisted of two steps: first, we asked the users to register and carry out some activities in the social learning platform and second, we invited them to answer a questionnaire. The recommendations thus have been made for each user based on her/his interactions data within the platform. The link to the questionnaire was only enabled when a user had already received recommendations. If a user had not received any recommendations yet, we showed a message "There is no recommendations for you today". This way, the users were able to explore their recommendations first and later to respond to the questionnaire based on their experience with their recommendations. In the rest of this section, we first describe the design of the user-centric evaluation and then, we present the questionnaire we used in our user study.

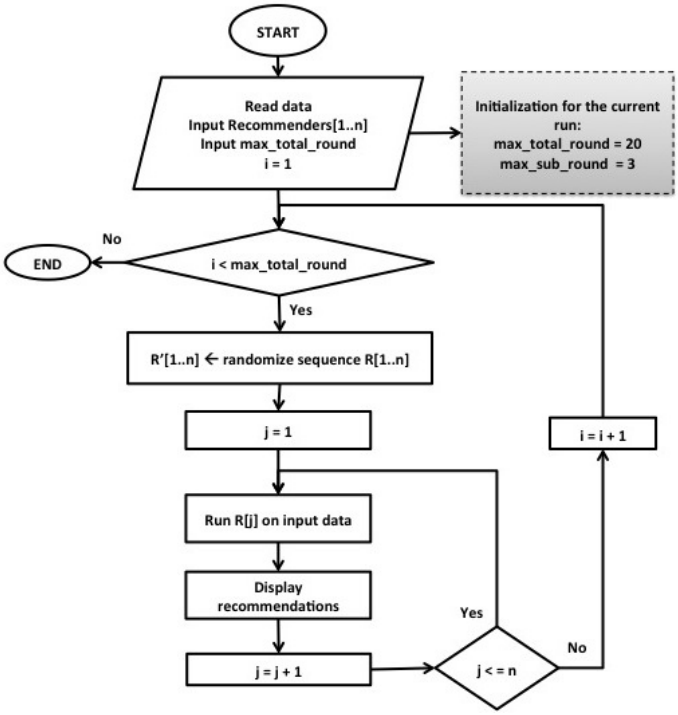


Fig. 1: Design of the user-centric evaluation with random sequence of running algorithms

2.4.1 Design

In principle, two types of tests are possible, a repeated measures design in which all users are tested repeatedly, once for each recommender; and a design in which each user is exposed to one recommender only, once and only once. In the first case, users are tested repeatedly (within-subjects design), in the second case users act as each others' replications (between-subjects design). Since it is impossible to guarantee that all users are indeed exposed to all recommenders and a repeated measures design with missing values is hard to analyse, only the second option is a feasible one. Besides, there is little a priori reason to expect that users have inherently different levels of responding (if that were the case, a repeated measures design is preferable as it removes variation due to those differences). Figure 1 shows the method used in the user-centric evaluation. We have a set of candidate recommender systems R1... Rn where n is number of candidate recommender algorithms. In this study, n equals three (3) since we selected three recommender algorithms: 1. a memory-based CF, 2. a model-based CF and 3. a graph-based CF.

Users will typically enter the ODS platform, be confronted with a recommendation list made either by R1, R2, or R3. They then are requested to answer the questionnaire; the questionnaire becomes available by clicking on a link provided. This means that there may be sequence effects, since participants enter in the experiment one after the other. To avoid such effects, treatments (types of recommendations) were assigned in a random order over time.

Since it is technically not feasible to administer a randomly drawn treatment per user recommendation event, treatments were administered in blocks of fixed time periods (randomized block design): R1-R2-R3, then R3-R1-R2, then R2-R1-R3. If there is a sequential effect, it will thus be balanced out over time. Since any one of the recommenders was active for all ODS users during the fixed time period it was tested, including those users who had already participated in the experiment, the questionnaire link was hidden from the latter to prevent them from participating multiple times.

2.4.2 Questionnaire

The questionnaire was designed to reflect how actual users perceive and appreciate the recommendations they receive, taking into account important aspects in user perception when running recommender systems' user studies [5], [28]. We asked the participants to answer six short questions by expressing their level of agreement with each of the questions. Agreement ranges from completely disagree (1) to completely agree (5). The questionnaire contains six statements: five questions regarding quality of the recommendations and one regarding the language of the recommendations. This is a rather low number, but we feared that the response rate would drop dramatically if we were to add more items: a recommendation is something to inspect immediately not after answering a lengthy questionnaire first. The description of the quality metrics were embedded with each question itself. Moreover, we added an open question at the end of the questionnaire through which the users can provide their general comments. The statements were:

1. The recommendations are relevant to my activities (Accuracy).
2. The recommendations provide me with novel information (Novelty).
3. The recommendations differ significantly from each other (Diversity).
4. The recommendations are useful for me (Usefulness).
5. The recommendations are surprising to me (Serendipity).
6. I am satisfied with the language of the recommendations.

For selecting these five quality metrics, we followed the ResQue framework presented by Pu et al. [29]. The framework provides a unified method for user-centric evaluations. However, making use of the whole framework can be very time consuming for participants since it includes many metrics. Therefore, we only focus on five important metrics that have been identified in the literature on recommender systems user studies as indicators of users satisfaction on the recommendations made for them [3], [29], [30]. By tending towards simplicity we seek to guarantee responsiveness.

In total, we had sixty participants from fifteen European countries: Greece, the Netherlands, Romania, the UK, Cyprus, Germany, Serbia, Bulgaria, Croatia, Estonia, Ireland, Lithuania, Poland, Portugal, and Spain. Figure 2 shows the distribution of the participants over these countries. In total, 48% of the participants were female and 52% were male. The participants were both primary and secondary school teachers, educational designers, educational advisors and content experts. The participants were randomly provided with recommendations based on three candidate algorithms: the graph-based method, the nearest-neighbours

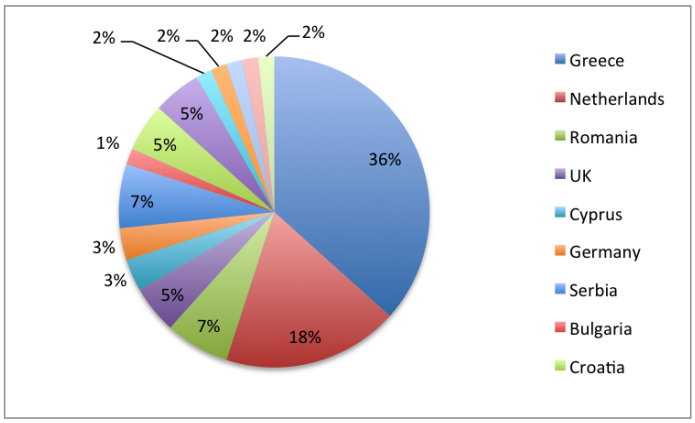


Fig. 2: Distribution of the participants over countries.

method (user-based kNN), and the matrix factorization method (BPRMF). We managed to obtain the same number of participants for the three algorithms, twenty for each. For user-centric evaluations in the recommender systems area, it has been claimed that "at least twenty (20) users" per condition is adequate to make a user-centric evaluation statistically sound [5].

2.5 Results

In this section, we first provide results of a traditional evaluation on the ODS data and then we present the user-centric evaluation results.

2.5.1 Data-centric Evaluation Results

The results of this offline data-centric evaluation on ODS data provide insights into the prediction accuracy of the recommendations made for ODS users. We conducted this offline evaluation in two steps, according to types of the CFs (memory-based or model-based):

Step 1: Evaluating three candidate memory-based CFs: the user-based graph-based approach, the user-based and item-based k-Nearest Neighbours methods (UserKNN and ItemKNN, respectively).

Step 2: Comparing performance of the candidate model-based CF that is a matrix factorization method (BPRMF) with the outperforming memory-based CFs from step 1.

Figure 3 shows results of step 1 that present precision and recall of memory-based CFs. For each memory-based CF algorithm, we evaluated five different sizes of neighborhoods (k=5,10,20,50,100). The horizontal axis (x) of both Figures 3(a) and Figure 3(b) indicate different sizes of neighborhood (k). The vertical axis (y) in Figure 3(a) and Figure 3(b) represent the values of precision and recall, respectively, at different cut-off values N (@N).

As Figure 3(b) shows, while the precision values of user-based CFs (UserKNN and Graph-based) improve by increasing size of neighbourhood (k), precision of item-based KNN (ItemKNN) declines by increasing the size of k. However, increasing the cut-off value (N) of precision from N=5 to N=10 improves the precision of ItemKNN, whereas precision of the user-based CFs (UserKNN and graph-based) decreases while N increases.

In general, UserKNN's precision@5 provides the highest values for precision from 0,047 (k=5) to 0.074 (k=100). The graph-based CF comes second place with precision@5 values increasing from 0.035 (k=5) to 0.060 (k=100). The highest value of precision for ItemKNN's is 0.026 (precision@5; k=5), which declines to 0.013 (precision@5; k=10).

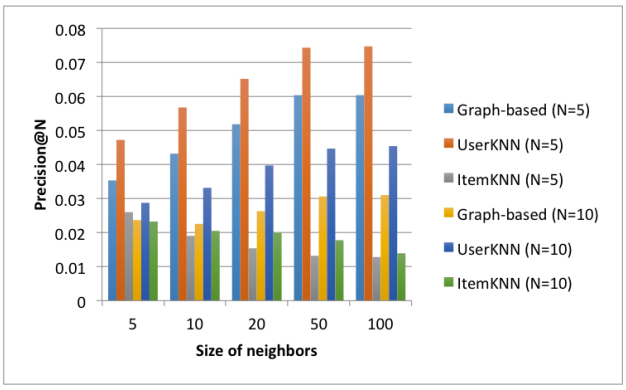
Similar to precision results, Figure 4b shows for recall that both user-based CFs (Graph-based and UserKNN) perform better than the item-based one (ItemKNN). In general, recall values for all algorithms increase when N increases from N=5 to N=10, which is expected in offline recommender system studies (Herlocker et al., 2004). The recall of the UserKNN and the graph-based CF changes for different neighbourhood sizes: for N=10, UserKNN's recall increases from 0.162 (k=5) to 0.283 (k=100) and the graph-based CF's recall increases from 0.166 (k=5) to 0.291 (k=100). The recall@10 for the ItemKNN goes from 0.1533 (k=5) to 0.0963 (k=10).

For the memory-based CFs, we set the size of neighbourhood (k) to 10. Although performance of the algorithms improves by increasing k in terms of accuracy metrics, we had to keep the neighbourhood size fairly small for reducing memory usage and also for making the recommendations generation task sufficiently fast for the user online evaluation. In a summary from step 1, we choose the graph-based and UserKNN CFs as the memory-based candidate CFs to be compared to the model-based matrix factorization method in the second step of the data-centric evaluation.

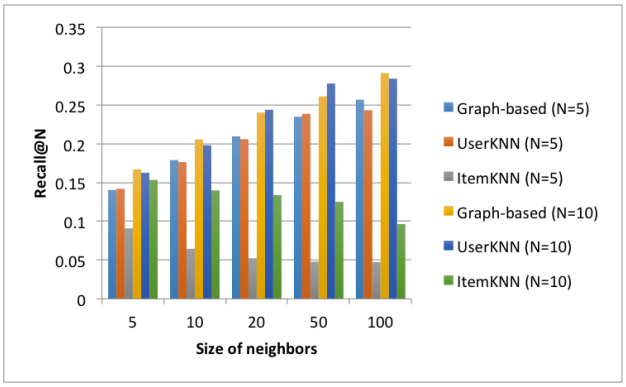
Figure 4 presents results of step 2 as a final comparison of different outperforming memory-based CFs (graph-based and UserKNN) with the candidate model-based matrix factorization method (BPRMF). For the model-based algorithm, we tried three different numbers of latent factors (3, 5, and 8). Among these three latent factors, BPRMF with f=8 achieved the best values for both precision and recall; consequently, we chose a number of latent factors equal to 8 for this final comparison. We set the learning rate (alpha) at 0.05 and the regularization parameter for user factors at 0.0025. The parameters have been tuned by using a validation set. The horizontal axis (x) in Figure 5 indicates the performance metrics in terms of precision and recall at two different cut-offs (N=5 and N=10). The vertical axis (y) shows values of precision@5, precision@10, recall@5, and recall@10 for different algorithms.

As Figure 4 shows, the user-based CFs (UserKNN and graph-based) outperform the matrix factorization method (BPRMF). The highest precision of BPRMF is precision@5=0.0135 whereas the lowest precision value for the user-based CFs is 0.0331 for the UserKNN's precision@10. For recall, the highest value for BPRMF (recall@10=0.0754) is still much smaller than the lowest recall@10 value for the memory-based CFs (UserKNN' s recall@5=0.1762).

In summary, the data-centric evaluation used in this study shows that the user-based CFs outperform the model-based CFs. According to conventional recommender systems evaluations, data scientists would use the user-based CFs algorithms in the live system as they outperformed the other candidate algorithms in the data-centric experiment. Since we want to investigate whether the user satisfaction results confirm the data-centric evaluation results, we apply the three candidate algorithms from both categories of model-based and memory-based CFs: the UserKNN, graph-based



(a) Precision@N of memory-based CFs for different sizes of neighbourhoods



(b) Recall@N of memory-based CFs for different sizes of neighbourhoods

Fig. 3: Comparison of memory-based CFs. Precision and recall scores (range: 0-1) for different sizes of neighbourhoods and for two cut-off points, N=5 and N=10.

CF and BPRMF in the user-centric evaluation part of the study.

2.5.2 *User-centric Evaluation Results*

Figure 5 shows percentage of answers in terms of level of agreement given by users on each of the five statements asked from users. The level of agreement ranges from 1 (completely disagree) to 5 (completely agree). Moreover, in contrast with the others, Figure 5(f) presents the average rating scores of each of the recommender algorithms for each of the five statements.

To analyse the results, we looked at each of the five quality metrics (five statements in the questionnaire). Each of the statements is mapped onto a quality metric, which each represents a dependent variable. The dependent variables are: 1. Usefulness, 2. Accuracy, 3. Novelty, 4. Diversity, and 5. Serendipity. We have one independent variable at three levels, corresponding to the three groups that are the recommender algorithms we used: 1. UserKNN, 2. Graph-based method, and 3. BPRMF. These three recommender algorithms have been selected based on the data-centric evaluation presented in previous section. For the sake of simplicity, from now on, we refer to "UserKNN" as "KNN" and to "BPRMF" as "MF"; thus, we have three experimental groups: 1. KNN, 2. Graph-based, and 3. MF.
As for the statistical test, we carried out five non-parametric univariate tests, one for each dependent variable (metric). We used Kruskal and Wallis (K-W). Note that in the literature the power of a KW test is found not to be much less than that of a parametric ANOVA (assuming the use of the latter is warranted) [31]. Since our new procedure now amounts making multiple comparisons by repeatedly testing the same subject - once for each metric - it is necessary to correct for the family-wise error rate. Therefore, we used a Bonferroni-Holm (B-H) correction.

Table 6 provides the results of the K-W test in an order of p-values magnitude for the three independent variables (1. KNN, 2. Graph-based method, and . MF) and the five dependent variables (1. Usefulness, 2. Accuracy, 3. Novelty, 4. Diversity, and 5. Serendipity). The results show that the algorithms are different in terms of usefulness, according to the K-W test for the variable usefulness p-value= 0.17 that seems to be significant (¡0.5). However, after applying B-H correction, there is no significance.
Furthermore, to be able to generalize over metrics and compare algorithms, we carried out a posteriori comparisons of medians and average ranks, using adjusted values of alpha.

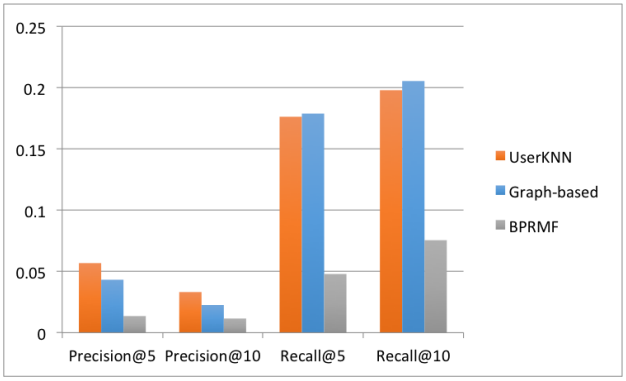
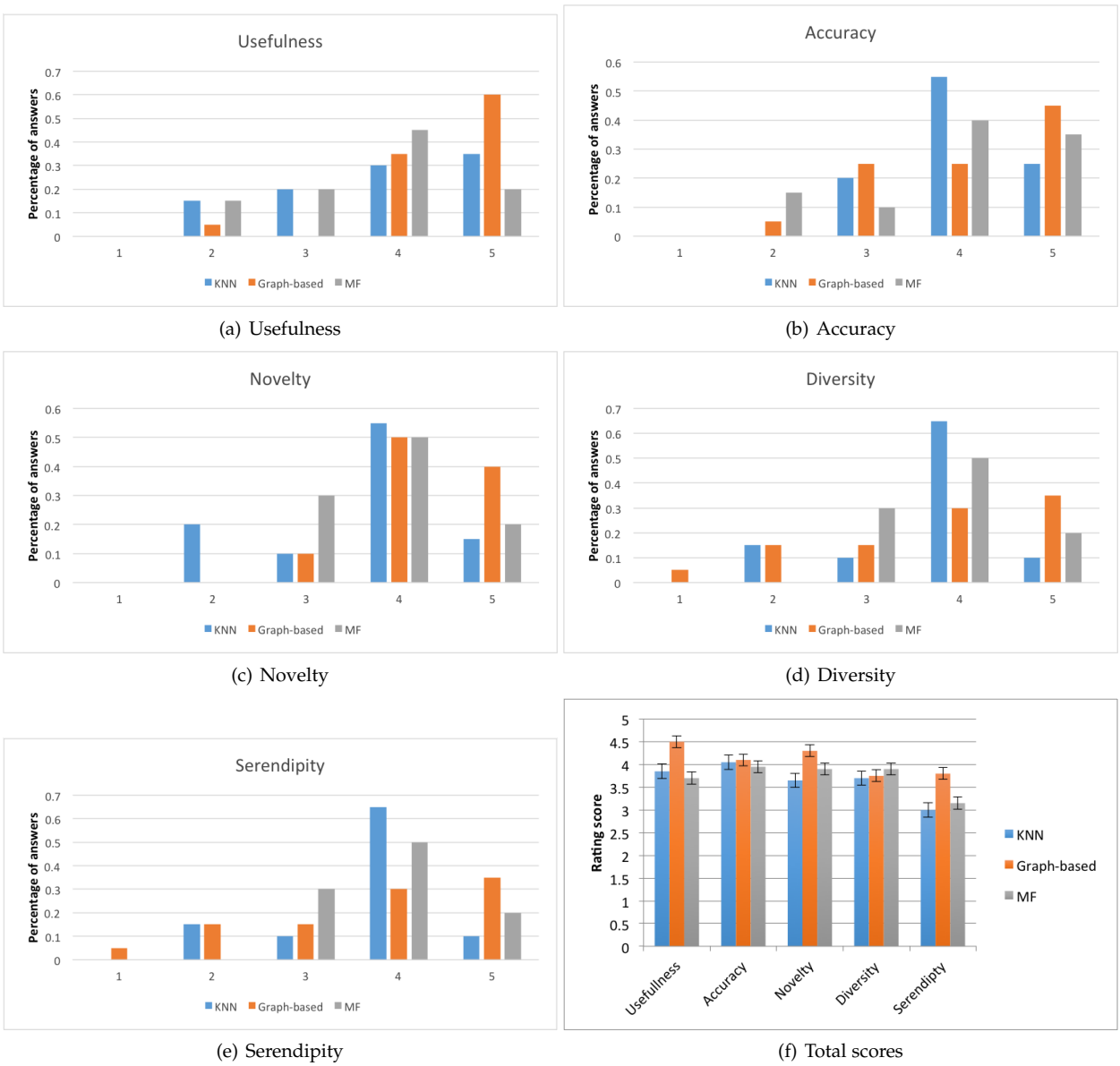


Fig. 4: Final comparison of the candidate CFs. Precision and recall scores (range: 0-1) for two cut-off points, N=5 and N=10.

3 Discussion and Conclusion

The main research question in this study is:
RQ: In social learning platforms, how is user satisfaction with recommender systems related to the performance of such systems measured in terms of their accuracy?
Our traditional, data-centric evaluation results (Figure 3 and Figure 4) show that the user-based nearest neighbors method outperforms other algorithms in terms of precision. As for recall, the nearest neighbours method and the graph-based method perform similarly and they both perform better than the matrix factorization method. However, the user-centric evaluation results show a quite different image. Based on the user-centric evaluation results (Figure 5), all three algorithms are not significantly different from a user's perspectives in terms of accuracy of the recommendations. In fact, users provide rather high average rating scores to all the algorithms (KNN: 4.05; graph-based: 4.1; MF: 3.95; all out of 5). Since our sample size was sufficiently large



(a) Usefulness (b) Accuracy

(c) Novelty (d) Diversity

(e) Serendipity (f) Total scores

Fig. 5: Percentage of answers of the online user-centric evaluation for the five independent variables (based on) (figures 6a-6e); Total average ratings (range: 1-5) (figure 6.f); N=60.

Fig. 6: Kruskal-Wallis Test For Five Quality Metrics; significant P-values are marked with a star (*).

Variable	p-value
Usefulness	0.017*
Serendipity	0.079
Novelty	0.105
Diversity	0.852
Accuracy	0.917

at 20 users per algorithm [5], we suggest this to show that the users were satisfied with the accuracy of the recommendations, regardless of the type of algorithm that generated them.

Nevertheless, the user-centric evaluation results (Figure 5) show that the graph-based recommender received a somewhat greater average rating score for perceived usefulness, novelty and serendipity of the recommendations by users, compared to the other two algorithms. If indeed there is such a difference, it is probably due to the fact that the graph-based recommender uses graph-walking methods to discover novel neighbours. These novel neighbours might be useful sources of information for a target user but they have no direct relations yet since they had no items rated in common. The further neighbours discovered by the graph-based method can provide useful, novel or serendipitous recommendations for a target user since they share less similarity with the target user and even they might be somehow dissimilar to some extent. Similarly, the average rating scores of the matrix factorization method for novelty, diversity, and serendipity of the recommendations perceived by users appears to be greater than the ones for the nearest neighbours method, although the difference is not

significant. For diversity of the recommendations, matrix factorization has got a greater average rating score (3.9 out of 5) than the one for the graph-based method (3.75 out of 5). Figure 5 suggests that, unlike the traditional evaluation results, the algorithms ranking order changes depending on the quality metrics (five statements) perceived by users. Based on the evaluation outcomes, the user-centric evaluation results in this study certainly do not confirm the traditional offline evaluation results. The few other recommender system studies that also take a user-centric view, show a similar inconsistency, although different in their details, between traditional evaluations and user-centric evaluations [8], [29], [30].

In general, the results show that the user-centric evaluation results do not confirm results of the traditional evaluation. Therefore, we conclude that it is necessary to study recommender systems from user-centric perspectives although user-centric evaluations are often complicated and costly in terms of time and resources. However, our study shows that recommender systems steered only by data-driven success indicators might guide data scientist to a less effective road in terms of users satisfaction. The results of this study need to be confirmed within a longitudinal study that tracks user satisfaction in the long run. It should also take into account more users than the 60 people we questioned. Preferably and if at all experimentally feasible, more questions should be used to delineate the five user satisfaction constructs (usefulness, accuracy, novelty, diversity, serendipity). In our view, then, our experimental design may serve the recommender system community to gain more insights into the differences between data-centric and user-centric evaluations measures for recommender systems.

ACKNOWLEDGMENTS

This paper is part of a doctoral study funded by NELLL (the Netherlands Laboratory for Lifelong Learning) and the Open Discovery Space project. Open Discovery Space is funded by the European Union under the Information and Communication Technologies (ICT) theme of the 7th Framework Programme for R&D. This document does not represent the opinion of the European Union, and the European Union is not responsible for any use that might be made of its content. The work of Hendrik Drachsler has been supported by the EU project LACE FP7 Program.

REFERENCES

[1] N. Manouselis, H. Drachsler, K. Verbert, and E. Duval, *Recommender systems for learning*. Springer Science & Business Media, 2012.

[2] J. Vassileva, "Toward social learning environments," *Learning Technologies, IEEE Transactions on*, vol. 1, no. 4, pp. 199–214, 2008.

[3] B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, and C. Newell, "Explaining the user experience of recommender systems," *User Modeling and User-Adapted Interaction*, vol. 22, no. 4-5, pp. 441–504, 2012.

[4] S. M. McNee, J. Riedl, and J. A. Konstan, "Being accurate is not enough: how accuracy metrics have hurt recommender systems," in *CHI'06 extended abstracts on Human factors in computing systems*. ACM, 2006, pp. 1097–1101.

[5] B. P. Knijnenburg, M. C. Willemsen, and A. Kobsa, "A pragmatic procedure to support the user-centric evaluation of recommender systems," in *Proceedings of the fifth ACM conference on Recommender systems*. ACM, 2011, pp. 321–324.

[6] S. M. McNee, J. Riedl, and J. A. Konstan, "Making recommendations better: an analytic model for human-recommender interaction," in *CHI'06 extended abstracts on Human factors in computing systems*. ACM, 2006, pp. 1103–1108.

[7] F. Martin, "Top 10 lessons learned developing, deploying, and operating real-world recommender systems, industry keynote 3rd acm conference on recommender system 2009, new york," 2009.

[8] M. A. Chatti, S. Dakova, H. Thus, and U. Schroeder, "Tag-based collaborative filtering recommendation in personal learning environments," *Learning Technologies, IEEE Transactions on*, vol. 6, no. 4, pp. 337–349, 2013.

[9] M. Erdt, A. Fernandez, and C. Rensing, "Evaluating recommender systems for technology enhanced learning: A quantitative survey," *Learning Technologies, IEEE Transactions on*, vol. 8, no. 4, pp. 326–344, 2015.

[10] J. Liu, P. Dolan, and E. R. Pedersen, "Personalized news recommendation based on click behavior," in *Proceedings of the 15th international conference on Intelligent user interfaces*. ACM, 2010, pp. 31–40.

[11] M. Richardson, E. Dominowska, and R. Ragno, "Predicting clicks: estimating the click-through rate for new ads," in *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007, pp. 521–530.

[12] S. Schnitzer, C. Rensing, S. Schmidt, K. Borchert, M. Hirth, and P. Tran-Gia, "Demands on task recommendation in crowdsourcing platforms-the worker?s perspective," in *CrowdRec Workshop*, 2015.

[13] M. Allahbakhsh, B. Benatallah, A. Ignjatovic, H. R. Motahari-Nezhad, E. Bertino, and S. Dustdar, "Quality control in crowdsourcing systems: Issues and directions," *IEEE Internet Computing*, vol. 17, pp. 76–81, 2013.

[14] H.-C. Schmitz, M. Scheffel, M. Friedrich, M. Jahn, K. Niemann, and M. Wolpers, "Camera for ple," in *Learning in the Synergy of Multiple Disciplines*. Springer, 2009, pp. 507–520.

[15] S. Fazeli, B. Loni, H. Drachsler, and P. Sloep, "Which recommender system can best fit social learning platforms?" in *Open Learning and Teaching in Educational Communities*. Springer, 2014, pp. 84–97.

[16] J. L. Herlocker, J. A. Konstan, and J. Riedl, "Explaining collaborative filtering recommendations," in *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. ACM, 2000, pp. 241–250.

[17] J. B. Schafer, D. Frankowski, J. Herlocker, and S. Sen, "Collaborative filtering recommender systems," in *The adaptive web*. Springer, 2007, pp. 291–324.

[18] I. Pilászy and D. Tikk, "Recommending new movies: even a few ratings are more valuable than metadata," in *Proceedings of the third ACM conference on Recommender systems*. ACM, 2009, pp. 93–100.

[19] A. Bellog?n, P. Castells, and I. Cantador, "Neighbor selection and weighting in user-based collaborative filtering?: A performance prediction approach," *ACM Transactions on the Web, In Press*, vol. 1, pp. 76–81, 2014.

[20] F. Ricci, L. Rokach, B. Shapira, and P. Kantor, "Recommender systems handbook. recommender systems handbook," ISBN 978-0-387-85819-7. Springer Science+ Business Media, LLC, 2011, 1, Tech. Rep., 2011.

[21] K. Verbert, H. Drachsler, N. Manouselis, M. Wolpers, R. Vuorikari, and E. Duval, "Dataset-driven research for improving recommender systems for learning," in *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*. ACM, 2011, pp. 44–53.

[22] S. Fazeli, A. Zarghami, N. Dokoohaki, and M. Matskin, "Mechanizing social trust-aware recommenders with t-index augmented trustworthiness," in *Trust, Privacy and Security in Digital Business*. Springer, 2010, pp. 202–213.

[23] J. A. Golbeck, "Computing and applying trust in web-based social networks," 2005.

[24] H. Ma, H. Yang, M. R. Lyu, and I. King, "Sorec: social recommendation using probabilistic matrix factorization," in *Proceedings of the 17th ACM conference on Information and knowledge management*. ACM, 2008, pp. 931–940.

[25] P. Massa and P. Avesani, "Trust-aware recommender systems," in *Proceedings of the 2007 ACM conference on Recommender systems*. ACM, 2007, pp. 17–24.

[26] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "Bpr: Bayesian personalized ranking from implicit feedback," in *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*. AUAI Press, 2009, pp. 452–461.

[27] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Transactions on Information Systems (TOIS)*, vol. 22, no. 1, pp. 5–53, 2004.
[28] B. P. Knijnenburg and M. C. Willemsen, "Evaluating recommender systems with user experiments," in *Recommender Systems Handbook*. Springer, 2015, pp. 309–352.
[29] P. Pu, L. Chen, and R. Hu, "A user-centric evaluation framework for recommender systems," in *Proceedings of the fifth ACM conference on Recommender systems*. ACM, 2011, pp. 157–164.
[30] A. Said, B. Fields, B. J. Jain, and S. Albayrak, "User-centric evaluation of a k-furthest neighbor collaborative filtering recommender algorithm," in *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 2013, pp. 1399–1408.
[31] A. Field, J. Miles, and Z. Field, "Discovering statistics using r. 2012."

Soude Fazeli is a postdoctoral researcher at Delft University of Technology, the Netherlands and a PhD candidate at Open University of the Netherlands. She received her master's degree in Software Engineering of Distributed Systems from Royal Institute of Technology (KTH), Sweden in 2009. Since her master's degree, she has been working on recommender systems' implementation, integration, and evaluation. She reviews papers for several journals and conferences in the field of computer science and also technology enhanced learning. She's currently working within the EU FP7 CrowdRec project (http://crowdrec.eu). The main focus of CrowdRec project is on next generation recommendations, which are real-time, large-scale, socially informed, interactive, and context aware.

Hendrik Drachsler is Associate Professor for Personalised Learning Technologies at the Welten Institute of the Open University of the Netherlands. His research interests include learning analytics, personalisation technologies, recommender systems, educational data, mobile devices, and their applications in the fields of technology-enhanced learning and health2.0. He is chairing the EATEL SIG dataTEL and the national SIG Learning Analytics of the Dutch umbrella organisation SURF. He is elected member of the Society of Learning Analytics Research (SoLAR). In the past he has been principal investigator and scientific coordinator of various national and EU projects (e.g., laceproject.eu, patient-project.eu, LinkedUp-project.eu). He regularly chairs international scientific events and is Associate Editor of IEEE's Transactions on Learning Technologies, and the Journal of Learning Analytics.

Marlies Bitter-Rijpkema works as assistant professor at the Welten Institute of the Open University of the Netherlands. She is fellow of ICO and SIKS and expert of the EADTU Empower network. She holds a PhD in Educational Technology dealing with virtual multidisciplinary teamwork. Over the years she worked in various roles as researcher, project manager and developer across national and projects and was liaison officer of the OU to the Dutch Digital University Consortium. Her current research focusing on co-creativity, organizational contexts, professional learning in organizational contexts ,social (open) innovation networks and business modeling materializes in the development of networks for innovation and learning on f.e.media literacy for social change in Asia (Medlit), supportive learning networks and training for female entrepreneurship (Digifem) and new librarianship (LibrarySchool, Biebkracht) as well as new learning support for beta sciences (NILMRT on Microreactortechnology, gamified (m-)learning for chemistry and civil engineering) and open education (recommenders in ODS, Open Scout).

Francis Brouns Francis Brouns has been employed at the Welten Institute since 1999 and has been working at the design, development and implementation of learning specifications and innovative learning environments in support of lifelong learning. In her position of Assistant Professor her research has evolved into supporting social and networked learning through technology enhanced learning, , embracing current developments in MOOCs and learning analytics. She takes part in various EU funded projects such as TENCompetence, LTfLL, and ODS. Currently she is involved in the EU CIP projects ECO and EMMA, both project considering MOOCs, developing pedagogical framework and learning analytics approaches.

Wim van der Vegt is a senior developer at the Welten Institute of the Open University of the Netherlands. He received his PhD from Open University of the Netherlands.

Peter B. Sloep (PhD) is full professor of Technology Enhanced Learning at the Welten Institute of the Open University of the Netherlands. He is honorary professor at the Caledonian Academy of Glasgow Caledonian University. His research encompasses such topics as networked learning (specifically but not exclusively for professionals), learning design, open educational resources, learning objects, standards for learning technologies, as well as knowledge sharing and creative collaboration in communities and networks. He co-authored more than 200 peer-reviewed publications in scholarly journals and conference proceedings, and has co-authored or edited three books. Sloep is a frequent speaker at national and international conferences. He frequently reviews papers for various journals and conferences in the TEL field.